

### A Dog's Breakfast?

Stephen J. O'Brien and William J. Murphy

Once the human genome sequence was drafted and later finished, it became evident that the ultimate rationale for the International Human Genome Project was not deposition of a generic 3 billion-nucleotide list into GenBank. The real goals are much broader: first, to interpret the meaning of the human genome sequence by annotation, and second, to apply the derived knowledge for the betterment of humankind. Interpretation and annotation derive from comparative genomics in the same spirit that human surgical procedures emerged from comparative anatomy; applications emerge from studies of the genetic influences affecting human susceptibility to disease. Comparative genomics for mammals, which advanced with the completion of draft genome sequences for mouse and rat, is further expanded by Kirkness and colleagues (1) on page 1898 of this issue. Here, they present a 1.5× whole-genome sequence of the domestic dog. With the dog sequence, genomics ventures from the laboratory haven of traditional rodent models of disease into the living room with an annotated whole-genome sequence of man's best friend.

Selecting the dog for full-genome sequencing was a sound decision, made both by Kirkness and colleagues and by GRASP, the committee of the National Human Genome Research Institute (NHGRI) that directs NIH dollars toward whole-genome sequencing. Dogs have 78 chromosomes, considerably reshuffled relative to the most recent common ancestor of Carnivora, the mammalian order in which canids reside (2). Domesticated since the dawn of agriculture 10,000 to 15,000 years ago, dogs today exist in more than 400 distinct breeds with a panoply of morphological and behavioral genetic differences suitable for genetic exploration (3). Dogs enjoy a medical surveillance and clinical literature second only to humans, succumbing to 360 genetic diseases that have human counterparts (4). Dogs have been beneficial for standard pharmaceutical safety assessment and also for groundbreaking gene therapy successes (5, 6). The progress in dog genetics has been astounding. A mere 4 years ago, when comparative mammalian ge-

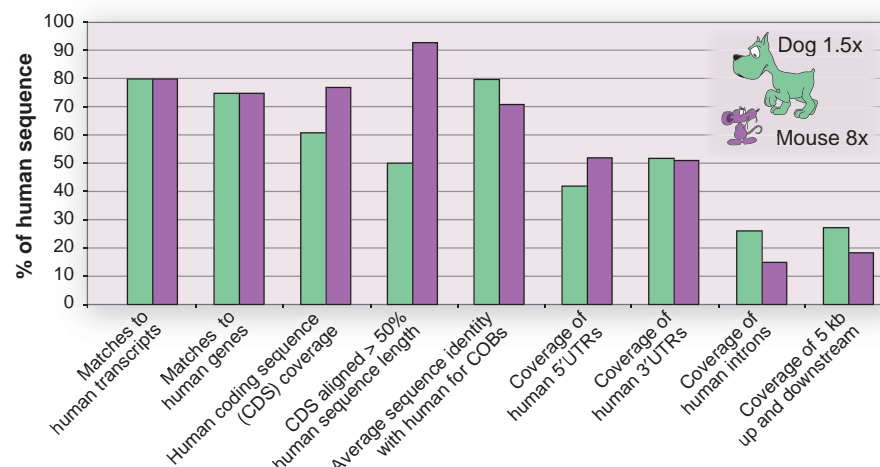
nomics was featured in the annual Genome Issue of *Science*, the dog genome was not included (7) because several linkage/syntenic groups were yet to be chromosomally assigned. Last year, several international conferences discussed dog genomics, and today we have a 1.5× genome sequence.

How was the dog genome sequence built? A whole shotgun sequence of 6.22 million reads was assembled into 1.9 million contiguous sequences (contigs) and 850,000 single sequence reads (singletons), which were connected into 522,101 scaffolds. Because 1.5× coverage covers about 78% of a complete mammalian genome, it is not possible to extend across chromosome-length distances without knowledge of the dog's physical or genetic map. Fortunately, Guyon *et al.* (8) have provided a canine radiation hybrid (RH) map of 3270 markers that allowed anchoring of sequence scaffolds to canine chromosome positions. To place contigs and scaffolds into their constituent dog chromosomes, the investigators relied on 2704 markers ordered on the canine RH map that could be stringently aligned to dog contigs or scaffolds. This resulted in a set of 2177 and 1766 "extended" dog markers mapped to the human and mouse genomes, respectively. Several criteria were applied to determine 159 conserved ordered syntenies between dog and human and 202 between dog and mouse.

What did they find? Actually, quite a bit,

including 18,473 orthologs of the 24,567 annotated human genes. The dog genome is small: 2.4 billion base pairs (bp) versus 2.9 billion bp for the human genome (9, 10). The primary reason for this difference is that humans have more repetitive DNA (46% of the genome versus 31% for dog and 38% for mouse). One of the most important discoveries is that even though the dog genome was surveyed at relatively low coverage (1.5×), it annotates slightly more human transcripts (29,563) and genes (18,473) than does the more complete 8× mouse sequence (29,529 transcripts, 18,311 genes). Much of this success may be due to the decreased nucleotide substitution rate in dog and human versus mouse (1, 11).

A provocative finding emerging from mammalian whole-genome sequencing is the widespread occurrence of what we call "conserved sequence blocks" (CSBs)—stretches of unique DNA sequence with high homology among human, mouse, and dog. In comparing mouse and human sequences, Mural *et al.* (12) called them "syntenic anchors" and Waterston *et al.* (11) called them "orthologous landmarks." Referring to them as "clusters of orthologous bases" (COBs), Kirkness *et al.* (1) define them as short sequences in which all three pairwise alignments are mutually best matches. Waterston *et al.* estimate 558,000 CSBs between human and mouse of median length 500 bp, for a total of 188 Mb (or 7.5% of the mouse genome). Kirkness *et al.* count a total of 371,774 CSBs of median length 456 bp in a human-mouse-dog genome comparison (total dog length = 169.4 Mb, or 7.0% of the dog genome). A dazzling 45% of the three species' CSBs are not associated with any genes or gene neighborhoods. In fact, 1.9%



**The genome of man's best friend.** Comparison of the human genome with the 1.5× canine genome (1) and the 8× mouse genome (11). UTRs, untranslated regions.

The authors are at the Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD 21702, USA.

of the dog genome is coding sequence, but more than 4% of the intergenic sequence is highly conserved between dog and human as a result of CSBs.

So what are these intergenic CSBs? At first glance, sequence conservation (reminiscent of evolutionary constraint on coding exons) might imply functional constraints on noncoding sequence that is the target of transcription factors, contains other regulatory elements, or influences higher order chromatin structure. However, it is also possible that some conserved regions simply accumulate mutations more slowly. The true explanation for these puzzling CSBs is a mystery, but the new dog data offer several insights. Mouse and human evolutionary lineages share a monophyletic ancestry in one of four superordinal lineages of placental mammals (the Eurochontoglires) (13). In contrast, dogs evolved in another superorder, Laurasiatheria, a finding now solidified by Kirkness *et al.* (1) and others (14) through phylogenetic analysis of repeat elements. Nonetheless, mice have an overall genome mutation rate that is twice that of humans and dogs, as affirmed by whole-genome comparisons (1, 11, 12). This mutation rate speedup for mice implies that, with respect to coding and intergenic regions, the similarity between the human and dog genomes is greater (by a factor of 2) than the similarity of the mouse genome to either dog or human. When CSBs are counted for dog versus human (643 Mb), they comprise nearly twice as many base pairs as mouse-human CSBs (374 Mb). Because one would predict that functional CSBs (as opposed to coincident nonfunctional CSBs) would be constrained like gene exons, their decay in mouse (twice the rate of that in human or dog) suggests that a fair portion may be constrained by nonfunctional forces.

Even if many of the approximately 400,000 CSBs are functionally inert, they still pose a huge bonus to the task of comparative genome alignment. They increase the number of comparative anchor sequences from 25,000 genes to 400,000 CSBs including the genes. By annotating the CSBs between any set of species, fine-scale syntenic order can be established with rather high precision.

Analysis of conserved dog-human synteny was achieved by determining the order of stretches of CSBs plus genes in conserved ordered synteny (CSOs) anchored by the dog RH map markers. Kirkness *et al.* discovered 159 CSOs between human and dog, analogous to the 342 CSOs reported between mouse and human (11). Human genome organization is highly conserved relative to the genomes of all placental mammals (15), and the mouse and dog genomes are highly rearranged through a larger proportion of

translocations than found in other mammalian genomes. A sizable portion of the 159 dog-human CSOs reflect chromosomal shuffles during the evolution of Canidae species over the last 30 to 50 million years, as at least 60 exchanges have occurred during the interval separating the carnivore ancestor from modern canids (2).

So how good is 1.5× dog genome coverage versus a more thorough 8× coverage for mouse? Actually, it is surprisingly good. The 1.5× dog sequence shows gene homologs for 80% of human transcripts and 78% of human genes, although only 83% of human transcripts were aligned for more than 50% of their length. Thus, although most dog counterparts of human genes were identified, the coding sequences were fragmentary. This could lead to errors in establishing orthology of noncoding regulatory sequences or identification of pseudogenes. The authors point out, however, that the 1.5× coverage provides a sufficient resource to go in and obtain full sequence of most coding regions fairly easily with targeted approaches.

Despite these caveats, the 1.5× dog sequence compares well with the 8× mouse sequence in coverage of genetically meaningful regions (see the figure). Further, because 4% of the dog intergenic sequence can be uniquely aligned with human orthologous sequences, this translates into twice as much homology between dog and human relative to mouse-human alignments. This alone portends well for future low-resolution (1× to 2×) genome sequences of mammals in other ordinal lineages.

The disadvantages of 1.5× coverage are listed by Kirkness *et al.* and include missing information (about 20%), incomplete sequence for most of the genes found, failure to resolve important segmental duplications [which constitute 5% of human and 1% of murine genomes (16)], and too many gaps for a chromosome length assembly. This latter issue is assuaged by constructing an RH map of a few thousand markers for index species. As for the dog assembly here, these RH map anchor loci can connect the contigs and scaffolds confidently to the genome organization of the new species by an integration of conserved CSB synteny to human and mouse genomes, bounded by the RH reference anchor markers.

The whole-genome dog sequence and the anticipated higher (6.5×) coverage projected by the NHGRI program—the public effort has already deposited 2.5× dog sequence in GenBank—will invigorate biological studies of dogs. This is the first mammalian species beyond laboratory rodents to enjoy a genome sequence. The dog sequence represents an additional clade of placental mammals with divergent evolutionary history and domestication-based

capture of genotypic diversity. But although the prospects are appealing, the dog model has its limits. First, because the dog (like the mouse) has an evolutionarily derived genome, the study of dog genome breakpoint junctions will tell us more about “dog-giness” than about humankind. Second, reproductive research on dogs is not advanced, which means that cloning, embryo transfer, stem cell research, transgenics, and gene knockout will be slow to develop. Third, ethical welfare concerns for companion animal research are greater than for rodents, limiting opportunities for experimental inquiry.

For mammals, the genome sequence derby is heating up. Human, mouse, rat, and now dog have produced appreciable genome sequence, with chimp, cow, and rhesus macaque scheduled for full-genome sequence by 2005 (17). NHGRI recently estimated that in the next 4 years, U.S. sequencing centers alone could produce 460 billion bases—the equivalent of 192 dog-sized genomes at 1× coverage (17). A small contingent of researchers (including ourselves) recently gathered for a workshop at MIT’s Whitehead Center for Genome Research to dream about which mammals to sequence next and at what coverage (1×, low; 3×, moderate; 7×, high). The bases for selection included representation of major superordinal clades across the mammalian radiation, biomedical relevance, annotation of the human genome at increasing scales of resolution, usefulness of reconstructing the patterns of genome reorganization, and the importance of capturing genomic inference in mammalian adaptation, development, and specialization (18). The recommendations converged on a group of 18 primate and 28 nonprimate species to be considered for whole-genome sequence analysis. Although costly, there is little doubt that comparative genomics of the mammalian radiations will greatly inform human biology as well as that of the 4600 species of mammals with which we share the planet.

## References

1. E. Kirkness *et al.*, *Science* **301**, 1898 (2003).
2. W. J. Nash *et al.*, *Cytogenet. Cell Genet.* **95**, 210 (2001).
3. E. A. Ostrander *et al.*, *Trends Genet.* **16**, 117 (2000).
4. D. F. Patterson, *Canine Genetic Disease Information System* (Mosby, St. Louis, MO, 2002).
5. G. M. Acland *et al.*, *Nature Genet.* **28**, 92 (2001).
6. K. P. Ponder *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13102 (2002).
7. S. J. O’Brien *et al.*, *Science* **286**, 458 (1999).
8. R. Guyon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5296 (2003).
9. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
10. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
11. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
12. R. J. Mural *et al.*, *Science* **296**, 1661 (2002).
13. W. J. Murphy *et al.*, *Science* **294**, 2348 (2001).
14. J. W. Thomas *et al.*, *Nature* **424**, 788 (2003).
15. W. J. Murphy *et al.*, *Genome Biol.* **2**, 0005.1 (2001).
16. E. E. Eichler, D. Sankoff, *Science* **301**, 793 (2003).
17. J. Couzin, *Science* **301**, 1176 (2003).
18. S. J. O’Brien *et al.*, *Science* **292**, 2264 (2001).